РАЗРАБОТКА ОБНОВЛЯЕМОГО ОБУЧАЮЩЕГО ДАТАСЕТА ДЛЯ НЕЙРОННОЙ СЕТИ АВТОМАТИЗИРОВАННОЙ ДИАГНОСТИКИ РАКА МОЛОЧНОЙ ЖЕЛЕЗЫ

А.Х. Оздиев[1], Д.А. Копцев, В.О. Веснина[2], И.Г. Фролова[2]
[1]Томский политехнический университет
[2]Научно-исследовательский институт онкологии, Томский национальный исследовательский медицинский центр Российской академии наук
ozdiev@tpu.ru

Введение

Смертность от злокачественных опухолей в России одна из самых высоких в мире, это можно объяснить отсутствием в нашей стране программ первичной профилактики и скрининга рака, т.е. своевременного выявления злокачественных опухолей. Большинство больных обращается с поздними стадиями заболевания, лечение которых оказывается малоуспешным. Главным препятствием внедрению современных методов профилактики в нашей стране является отсутствие научно обоснованной программы профилактики рака. Анализ причин снижения смертности от злокачественных опухолей в странах Европейского союза, США и Австралии показал, что снижение смертности произошло в результате успешных программ первичной профилактики и скрининга рака. В этом свете исследования, направленные на развитие технологий скрининга рака, являются как никогда актуальными.

Рак молочной железы занимает первое место среди онкологических заболеваний женщин – 16% всех случаев рака. По статистике, каждая 8-я женщина рискует получить диагноз РМЖ. В структуре смертности населения России от злокачественных новообразований раку молочной железы принадлежит 7,5% всех случаев [1]. Приведённая выше информация подтверждает актуальность поиска новых, более совершенных инструментов диагностики рака. С аппаратной точки зрения перспективными являются методы рентгеновской диагностики, основанные на применении фазового контраста и контраста рассеяния, для визуализации маммографических данных. С точки зрения программного обеспечения в последнее время на первый план выходит технология диагностики с применением методов машинного обучения. Однако, для работы в этом направлении необходимо создать базу маммографических снимков, пригодных для проведения экспериментальных работ с алгоритмами машинного обучения. Решению этой задачи и посвящена данная работа.

Источники данных и инструменты формирования датасета

Обучающий датасет сформирован из данных, полученных от двух источников. Первый – от партнеров из Научно-исследовательского института онкологии Томского национального исследовательского медицинского центра Российской академии наук. Вторая часть данных получена от исследовательской группы Breast Research Group из Hospital de São João, Breast Centre, Порту, Португалия [2]. В общей сложности собранный датасет состоит из 533 случаев, для каждого пациента из которых присутствует от 2 до 4 маммографических снимков.

Для предварительной обработки данных было использовано самостоятельно разработанное программное обеспечение в виде скриптов на языке Python, которые включали в себя инструменты для конвертации, сортировки, переименования, подписи изображений, а также обработки текстовых файлов с диагнозами.

Для разметки изображений было использовано бесплатное специализированное программное обеспечение (рисунок 1) под названием Colabeler AI Labeling Tool [3].

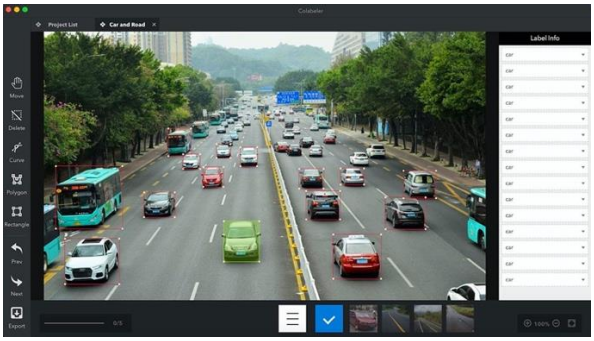

Рис. 1. Интерфейс Colabeler AI Labeling Tool

Программа позволяет формировать классы абстракций для разметки и непосредственно производить разметку изображений, предоставляю для этого необходимый набор инструментов. Области изображений содержащие размечаемые абстракции выделяются прямоугольниками или полигональными фигурами, вершины которых сохраняются в виде координат точек в системе координат текущего изображения. Для каждой выделенной области выбирается один из заданных классов абстракций.

Структура датасета

Разработанный датасет (на настоящий момент) представляет собой таблицу из 4 столбцов и 1978 строчек. Первый столбец содержит в себе маммографические проекции, которые имеют порядковый номер, обозначающий номер пациента, и порядковый номер проекции (рисунок

2-а). Второй столбец (рисунок 2-б) содержит в себе файл с описанием случаев и диагнозом на русском языке, третий столбец содержит аналогичный файл на английском языке (рисунок 2-в), а в четвертом – находится xml файл с разметкой изображения (рисунок 2-г). Схематичное изображение структуры строки представлено на рисунке 2.

а б в г

Рис. 2. Структура строки таблицы датасета

Структура файла разметки

XML файл разметки представляет собой набор тэгов, информация в которых и является результатом разметки:

- path – директория, где располагается данное изображение;
- outputs – размеченные объекты абстракций;
- name – назначенный размеченному объекту класс абстракций;
- bdnbox – размеченная область, содержащие координаты точек разметки;
- labeled – флаг, означающий наличие размеченных областей;
- size – размер данного изображения.

Тэг bndbox содержит координаты точек, также заключенные в тэги: xmin, xmax, ymin, ymax. Если тэг labeled содержит «false» это означает, что данное изображение **было** проанализировано, но образований, подлежащих классификации и локализации не найдено. Для размеченных изображений выделено 4 абстракции:

- **benign** – доброкачественная опухоль;
- **cancer** – злокачестенная опухоль;
- **nncancer** – опухоли с неизвестной первичной локализацие;
- **notclassified** – невозможно классифировать, требуется дополнительный анализ.

Как уже упоминалось, изображение может не содержать ни один из представленных классов абстракций. Пример маммографической проекции представлен на рисунке 3.

Таким образом, в ситуациях, когда у пациентки присутствует патология с её ярко выраженными признаками и когда у пациентов отсутствуют какие-либо признаки патологии, модель обученная, распознавать объекты вышеприведенных классов, может позволить освободить врача-рентгенолога от необходимости анализировать большую часть данных, экономя его время, силы и позволяя сконцентрироваться на более сложных случаях.

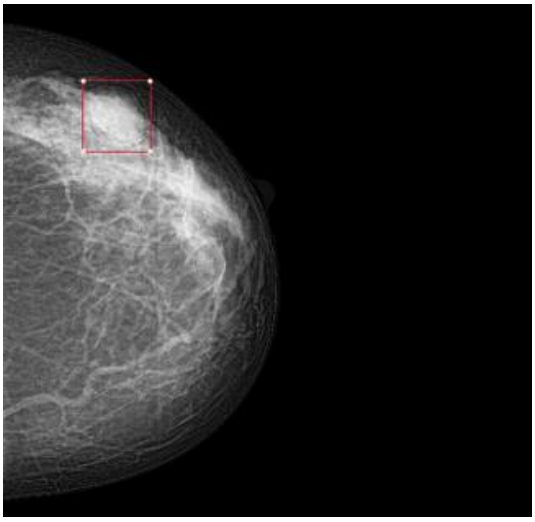

Рис. 3. Пример размеченного маммографического снимка

Заключение

В настоящий момент проблема смертности от раковых заболеваний находится на критичном уровне. Причиной этого является множество факторов, в том числе некачественная и несвоевременная диагностика. В случае рака молочной железы применение автоматизированных систем диагностике на основе машинного обучения и нейронных сетей может позволить решить несколько проблем. Во-первых, на местах, где нет квалифицированных специалистов для проведения диагностики по результатам маммографического сканирования, подобная технология может применяться для организации первичной диагностики, по результатам которой пациентки направлялись бы к специалисту для диагностики следующего уровня и уточнения диагноза.

Для разработки подобных инструментов диагностики необходимо создать большой набор достоверных данных для обучения и тренировки моделей. Данная работа направлена как раз на решение этой первостепенной задачи. Мы приглашаем к сотрудничеству научные коллективы, занимающиеся этими и смежными задачами.

Работа выполнена в рамках Программы повышения конкурентоспособности ТПУ, проект №ВИУ-ИШНКБ-62/2019.

Список использованных источников

1. А.Д. Каприн, В.В. Старинский, Г.В. Петрова, Злокачественные новообразования в России в 2018 году, Российский Центр информационных технологий и эпидемиологических исследований в области онкологии, Москва, 2019
2. http://medicalresearch.inescporto.pt/breastresearch/index.php/
3. http://www.colabeler.com/